Joint estimation of multiple networks from time course data

Chris. J. Oates*and S. Mukherjee†

February 11, 2013

Abstract

Graphical models are widely used to make inferences concerning interplay in multivariate systems, as modeled by a conditional independence graph or network. In many applications, data are collected from multiple individuals whose networks may differ but are likely to share many features. Here we present a hierarchical Bayesian formulation for joint estimation of such networks. The formulation is general and can be applied to a number of specific graphical models. Motivated by applications in biology, we focus on time-course data with interventions and introduce a computationally efficient, deterministic algorithm for exact inference in this setting. Application of the proposed method to simulated data demonstrates that joint estimation can improve ability to infer individual networks as well as differences between them. Finally, we describe approximations which are still more computationally efficient than the exact algorithm and demonstrate good empirical performance.

1 Introduction

Graphical models are widely used to model multivariate systems. Estimation of conditional independence structure (often called "network inference" or "structure learning") is increasingly a mainstream approach, for example in computational biology. Given data $\mathbf{X}$ a network estimator gives an estimate $\hat{G}(\mathbf{X})$ of the conditional independence graph G. The type of graph and its scientific interpretation depend on the model and scientific context.

In many applications, data is collected on multiple individuals $j \in \mathcal{J}$ that may differ with respect to interplay between variables, such that corresponding conditional independence graphs G^j may be individual-specific. For example, in biology, individuals may correspond to different patients or cell lines and the networks themselves to gene regulatory or protein signaling networks. Interplay in such networks can depend on the genetic and epigenetic state of the individuals, such that even for a well-defined system, such as signaling downstream of a certain receptor class, or a sub-part of the transcriptional program, details may differ between even closely related samples (Ideker and Krogan, 2012). For example, in yeast signaling, edges in the well-understood mitogen-activated protein kinase (MAPK) pathway can change depending on context (Zalatan *et al.*, 2012), whilst in cancer, it is thought that individual cell lines may differ with respect to signaling network connections. Continuing reduction in the unit cost of biochemical assays has led to an increase in experimental designs that include panels of potentially heterogeneous individuals (Barretina *et al.*, 2012; Cao *et al.*, 2011; Maher, 2012; The Cancer Genome Atlas Network, 2012). In such settings, given individual specific data $\mathbf{y}^j$, there is scientific interest in the individual specific networks G^j and their similarities and differences.

The case of multiple related individuals poses a number of statistical challenges for network inference:

- **Efficiency.** If the networks share features, then individual-level estimation (i.e. $\hat{G}^j = \hat{G}(\mathbf{y}^j)$), may be inefficient, since there is no sharing of information at the population level. Although individual network estimators $\hat{G}^j$ may be well-behaved as the individual-specific sample size n_j grows large (e.g.

[*] Centre for Complexity Science, University of Warwick, Coventry, UK.
[†] Department of Biochemistry, Netherlands Cancer Institute, Amsterdam, The Netherlands.

Kalisch and Bühlmann (2007)), in practice small-to-moderate n_j's and the inherently high-dimensional nature of network inference render inference challenging.

- **Data aggregation.** Aggregating data from multiple individuals and then performing network inference offers a way to obtain larger sample sizes. However, in settings where data from individuals are inhomogeneous (in the sense that the graphs G^j may differ between individuals), inferences regarding conditional independence structure cannot in general be obtained from aggregated data (Simpson's paradox) and testing whether data aggregation is appropriate may be challenging (Pearl, 1998). Estimating sufficiently homogeneous groups using mixture models and related clustering approaches offers an alternative (Zhou *et al.*, 2009; Mukherjee and Hill, 2011; Rodríguez *et al.*, 2011; Vu *et al.*, 2012; Hill and Mukherjee, 2013), but is challenging in the network setting, as we discuss further below.

- **Ancillary information.** Ancillary information may be available both at the "global" (population) and "local" (individual) levels. For example, in gene regulation, the biological literature provides general information concerning gene-gene interplay, whilst patient-specific characteristics might also be available. When such ancillary information is available it may be desirable to include it in inference (the "conditionality principle"), but doing so requires care in prior elicitation (Baumbach *et al.*, 2009; Wei and Pan, 2012).

In this paper we present a Bayesian approach to joint estimation of networks. The high-level formulation we propose is general and could be applied to a wide range of graphical model formulations. We present a detailed development for the time-course setting, focusing on directed graphical models called Dynamic Bayesian Networks (DBNs). These are directed acyclic graphs (DAGs) with explicit time-indices (Murphy, 2002). The main features of our approach are:

- **Bayesian framework.** We use a hierarchical Bayesian model, summarized in Fig. 2. Regularization is achieved using priors over both parameters and networks. We focus in particular on regularization of individual networks, introducing a latent network G to couple inference across the population. We report posterior marginal inclusion probabilities for every possible edge, thus providing a confidence measure for the inferred network topologies and offering robustness in settings where posterior mass is not highly concentrated on a single model.

- **Computationally efficient estimation from time-course data.** For the time-course setting, we put forward an efficient and deterministic algorithm. This is done by exploiting modularity of the DBN likelihood (Hill *et al.*, 2012) coupled with a sparsity restriction and a sum-product-type algorithm. In moderate-dimensional settings this allows exact joint estimation to be carried out in seconds to minutes (we discuss computational complexity below) making our approach suitable for interactive use.

- **Incorporation of ancillary information.** We allow for the inclusion of individual-specific ancillary information. Following Spencer and Mukherjee (2012) we also allow for interventional data, in which time courses are obtained under external intervention on network nodes.

Joint estimation of graphical models has recently been discussed in the penalized likelihood literature, with contributions including Danaher *et al.* (2012); Guo *et al.* (2011); Yang *et al.* (2012). In these studies, L_1 penalties, such as the fused graphical LASSO, are used to couple together inference of Gaussian graphical models (GGMs). Our work complements these efforts by offering a Bayesian formulation of joint estimation. This facilitates regularization using prior and ancillary network information. Moreover, our approach provides a natural way to estimate confidence in the inferred structure, providing robustness in multi-modal problems (Claassen and Heskes, 2012). Further, we focus on the time-course setting and DBNs rather than static data and GGMs. However, we note that unlike the above penalized approaches the Bayesian approach we propose is not well-suited to extremely high-dimensional settings with thousands of variables.

A recent paper by Penfold *et al.* (2012) considers Bayesian joint estimation for time-course data. Our work is in the same vein but differs in two main respects. First, we allow for prior information regarding the network structure and ancillary information including individual-specific characteristics. Network priors and ancillary information can usefully constrain inference, not least in biological settings. For example in the cancer signaling example we consider below, much is known concerning relevant biochemistry (Fig. 1) and individual-specific information pertaining to e.g. mutation status and receptor expression is often available (nowadays also in the clinical setting). Second, for the time-course setting, the exact algorithm we propose offers massive computational gains in comparison to the approach proposed by Penfold *et al.* (2012). As we discuss in detail below the methodology of Penfold *et al.* (2012) is prohibitively computationally expensive for the applications we consider here. Third, the computational efficiency of our approach allows us to present a much more extensive study of joint estimation, using both simulated and real data, than has hitherto been possible. This adds to our understanding of the performance of hierarchical Bayesian formulations for joint estimation.

Mixtures of graphical models have been used to explore heterogeneous populations (Zhou *et al.*, 2009; Mukherjee and Hill, 2011; Rodríguez *et al.*, 2011; Vu *et al.*, 2012; Hill and Mukherjee, 2013). However, mixture modeling requires the strong assumption that there exist groups which are (sufficiently) homogeneous with respect to model parameters. Otherwise, mixture components are forced to model heterogeneous populations, resulting in potentially poor fit and networks that may not be scientifically meaningful. Moreover, while graphical model estimation remains non-trivial, mixtures of graphical models are still more challenging, due to a number of factors relating to the hidden nature (and number) of the mixture components.

Further related work includes Werhli and Husmeier (2008), who propose a Bayesian approach to network inference based on multiple, steady-state datasets where in each dataset only a subset of the (shared) underlying network is identifiable. Dondelinger *et al.* (2012) extend the information sharing scheme from Werhli and Husmeier (2008) in the context of inference for time-varying networks. Hoff (2009) considers covariance estimation from a heterogeneous population, treating individual covariance matrices as samples from a matrix-valued probability distribution. Network priors have been discussed in the literature, including Imoto *et al.* (2003); Mukherjee and Speed (2008); Wei and Pan (2012). Our work differs from these efforts by focusing on joint estimation; as we describe below, this leads to a different model structure and prior specification.

The remainder of the paper is organized as follows. In Section 2 we lay out a hierarchical Bayesian formulation and in Section 3 we discuss computationally efficient exact inference. Empirical results are presented in Section 4 using simulated (Section 4) datasets. Finally we close with a discussion of our findings in Section 5.

2 Joint network inference

We carry out joint network inference using the hierarchical model shown in Fig. 2 that includes a prior network (G^0) as well as a latent network (G); each individual network (G^j; we use superscript notation when referring to a particular individual) is conceptually viewed as a variation upon the latter. Individual data $\boldsymbol{y}^j$ are then conditional upon individual networks. Estimates of the individual networks G^j are regularized by shrinkage towards the common latent network G which in turn may be constrained by an informative network prior. Since the latent network is itself estimated, this allows for adaptive regularization.

2.1 Hierarchical model

Consider the space $\mathcal{G}$ of (directed) networks (not necessarily acyclic) on the vertex set $\mathcal{P} = \{1, \ldots, P\}$. A network $G \in \mathcal{G}$ decomposes over parent sets as $G = G_1 \times \cdots \times G_P$ where $G_p \subseteq \mathcal{P}$ are the network parents of $p \in \mathcal{P}$. Write $\mathcal{G}_p$ for the set of possible parent sets for variable p, such that formally $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_P$. Write $\mathcal{J} = \{1, \ldots, J\}$ for the set of individuals in the population.

As shown in Fig. 2, each individual network G^j is conditional on a latent network G which in turn depends on a prior network G^0 (Section 2.2). As in any graphical model, data $\boldsymbol{y}^j$ is conditional on graph G^j and parameters $\boldsymbol{\theta}^j$; A^j denotes any ancillary information available on individual j. In this Section we describe our general model and network priors, while in Section 3 we discuss the special case of inference for

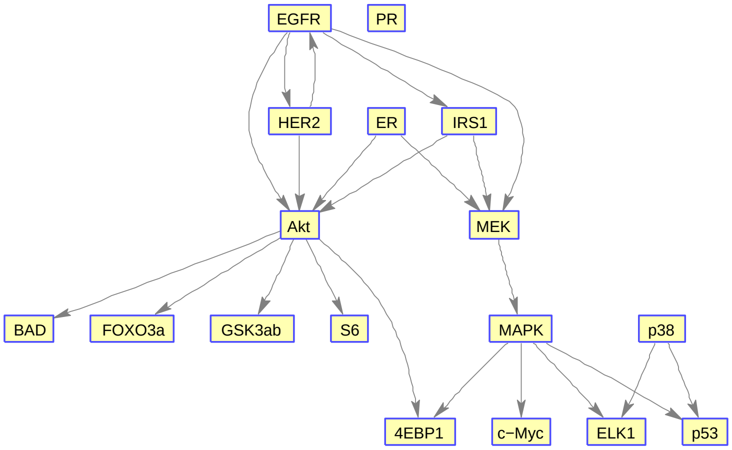


Figure 1: Epidermal growth factor receptor (EGFR) pathway for mammalian cells, characterized by extensive biochemistry. [Here edges represent high-level summaries of often complex molecular interactions that may involve latent chemical species.]

time-course data, giving full details of the likelihood for that case. The model is specified by

$$p(G|G^0,\eta) \quad \propto \quad \exp\left(-\eta d(G,G^0)\right) \tag{1}$$

$$p(G^1,\dots,G^J|G,\boldsymbol{\lambda},A^1,\dots,A^J) \quad \propto \quad \exp\left(-\sum_{j\in\mathcal{J}}\lambda^j d^j(G^j,G;A^j)\right) \tag{2}$$

where the functionals $d^j, d : \mathcal{G}\times\mathcal{G}\to\mathbb{R}$ and hyperparameters $\eta,\lambda^1,\dots,\lambda^J$ must be specified (Section 2.2). This formulation is borrowed from statistical mechanics, where d^j, d may be interpreted as energy terms, $\eta,\lambda^1,\dots,\lambda^J$ as inverse temperature parameters and Eqns. 1,2 as Boltzmann (or Gibbs) distributions. Taken together with a suitable graphical model likelihood $p(\boldsymbol{y}^j|G^j,\boldsymbol{\theta}^j)$, we obtain the data-generating model. JNI performs inference jointly over $(G,G^1,\dots,G^J)$, with information sharing occurring via the latent network G. The use of a latent network follows Guo *et al.* (2011); Imoto *et al,* (2006); Penfold *et al.* (2012); Werhli and Husmeier (2008). In some biological settings, it may be natural to think of the latent network as describing a "wild type network", however such an interpretation is not required. We refer to this general formulation as joint network inference (JNI).

2.2 Network prior

Specifying a network prior (Eqn. 1) requires a penalty functional $d:\mathcal{G}\times\mathcal{G}\to\mathbb{R}$ and a prior network $G^0\in\mathcal{G}$, with the former capturing how close a candidate network $G\in\mathcal{G}$ is to the latter (Imoto *et al.*, 2003; Mukherjee and Speed, 2008). We discuss choice of G^0 below. Given G^0, a simple choice of penalty function d is the structural Hamming distance $d(G,G^0)=\mathrm{SHD}(G,G^0):=\sum_{p\in\mathcal{P}}|G_p\Delta G^0_p|$. Here $A\Delta B$ denotes the symmetric difference of sets A and B and $|A|$ denotes cardinality of the set A. The hyperparameter η controls the strength of the prior network G^0 (Eqn. 1). For brevity we follow Penfold *et al.* (2012) by restricting attention to SHD priors, however our formulation is general (see below) and compatible with other penalty functionals. For their work on joint estimation of inverse covariance matrices, Danaher *et al.* (2012); Yang *et al.* (2012) employed the fused graphical LASSO (FGL) penalty, which may be interpreted

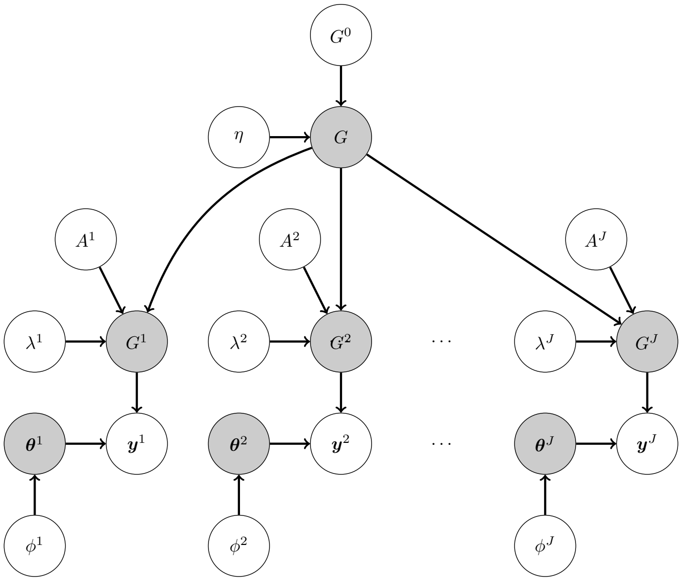


Figure 2: Joint network inference (JNI): A hierarchical model for analysis of multivariate data from a heterogeneous population. [Shaded nodes are unobserved. G^0 = prior network, G = latent network, G^j = individual j's network, $\boldsymbol{\theta}^j$ = parameters for individual j, $\boldsymbol{y}^j$ = data obtained on individual j, A^j = ancillary information available on individual j, η, λ^j = inverse temperature hyperparameters, ϕ^j = parameters defining a prior on $\boldsymbol{\theta}^j$.]

as a real-valued extension of SHD (strictly speaking, there is no analogue of the latent network G here; FGL directly penalizes the difference between individual networks G^j, G^k). Another interesting extension due to Werhli and Husmeier (2008) distinguishes $G^0 \setminus G$ ("false prior positives") and $G \setminus G^0$ ("false prior negatives") by allocating a separate inverse temperature hyperparameter for each case. Alternatively, one could employ a binomial prior as described in Dondelinger *et al.* (2012), which provides the same distinction, but allows for the hyperparameters of the binomial to be integrated out.

Conditional on a latent network G, individual networks G^j are regularized in a similar way, as $d^j(G^j, G) = \text{SHD}(G^j, G)$. In their work on combining multiple data sources, Werhli and Husmeier (2008) allow the λ^j to vary over individuals (data sources) $j \in \mathcal{J}$, with λ^j reflecting the quality of dataset j. Likewise Penfold *et al.* (2012) learn the λ^j on an individual by individual basis. However, in both studies, hyperparameter elicitation is non-trivial (see Section 2.4). To further limit scope, we consider only the special case where $\lambda^1 = \lambda^2 = \cdots = \lambda^J := \lambda$.

When ancillary information A^j is available regarding a specific individual network G^j, it is desirable to augment the prior specification in such a way as to condition upon A^j. In general such modification will be application specific.

Although we focus on SHD priors, the inference procedures presented in this paper apply to the more general class of modular priors, which may be written in the form

$$d(G, G^0) = \sum_{p \in \mathcal{P}} d_p(G_p, G_p^0), \quad d^j(G^j, G; A^j) = \sum_{p \in \mathcal{P}} d_p^j(G_p^j, G_p; A^j) \tag{3}$$

for some functionals $d_p, d_p^j : \mathcal{G}_p \times \mathcal{G}_p \to \mathbb{R}$. Modularity here refers to a factorization over variables $p \in \mathcal{P}$, implying that only local information is available *a priori*. The SHD priors are clearly modular.

2.3 Two special cases: INI and ANI

Up to inclusion of ancillary information, prior strength is fully determined, in this simplified setting, by the parameter pair (λ, η). Taking $\eta \to \infty$ requires that the latent network G is (almost surely) identical to the prior network G^0; in the limit this corresponds to treating network inference for each individual separately, i.e. the estimator $\hat{G}^j = \hat{G}(\mathbf{y}^j)$. We call this approach "independent network inference" (INI). Conversely, taking $\lambda \to \infty$ requires that (almost surely) individual networks G^j do not deviate from the latent network G; this corresponds to assuming individuals have identical (unknown) network structure, but allowing parameter values $\boldsymbol{\theta}^j$ to vary between individials, possibly becoming equal to zero. We call this approach "aggregated network inference" (ANI). Taking $\lambda, \eta \to \infty$ together corresponds to using only the prior. A further, cruder, approach would be to simply combine all data in order to estimate a single network and parameter set, an approach which Werhli and Husmeier (2008) call "monolithic". We compare these approaches empirically in Section 4.

2.4 Network prior elicitation

Elicitation of hyperparameters for network priors is an important and non-trivial issue. Hyperparameters can be set using the data, but this poses a number of challenges, as reported in Dondelinger *et al.* (2012); Penfold *et al.* (2012); Werhli and Husmeier (2008). In the context of sequential hierarchical network priors, Dondelinger *et al.* (2012) observed that when there is limited data available, hyperparameters inferred from the data may be biased towards imposing too much agreement with the prior. Penfold *et al.* (2012) used an improper hyperprior over the individual inverse temperature parameters λ^j, reporting that for most individuals posterior marginals did not differ greatly from the prior (possibly due to uninformative data). Similarly Werhli and Husmeier (2008) assigned improper flat prior distributions over the hyperparameters, reporting that estimation was rather difficult. Due to such weak identifiability of hyperparameters, we chose instead to specify the hyperparameters λ, η in a subjective manner.

For subjective elicitation of network hyperparameters, interpretable criteria are important. We present three criteria below which, for the special case of SHD which we consider, are simple to implement and can be used for expert elicitation. These heuristics seek to relate the hyperparameters to more directly interpretable measures of the similarity and difference which they induce between prior, latent and individual networks.

Firstly, we note the following formula for the probability of maintaining edge status (present/absent) between the latent network G and an individual network G^j:

$$h_\lambda := p(i \notin G_p^j \Delta G_p) = \frac{e^{-\lambda \times 0}}{e^{-\lambda \times 0} + e^{-\lambda \times 1}} = \frac{1}{1 + e^{-\lambda}}. \tag{4}$$

This probability provides an interpretable way to consider the influence of λ. For example a prior confidence of $h_\lambda \approx 0.73$ that a given edge status in G is preserved in a particular individual G^j translates into a hyperparameter $\lambda \approx 1$ (see SFig. 1). An analogous equation relates η and $h_\eta := p(i \notin G_p \Delta G_p^0)$, allowing prior strength to be set in terms of the probability that an edge status in the prior network G^0 is maintained in the latent network G.

A second, related approach is to consider the expected total SHD between an individual network G^j and the wild type network G:

$$\mathbb{E}(\mathrm{SHD}(G^j, G)) = P^2(1 - h_\lambda) \tag{5}$$

This can be interpreted as the average number of edge changes needed to obtain G^j from G. An analogous equation holds for η and h_η.

Thirdly, in certain applications, the latent network G may not have a direct scientific interpretation, in which case the criteria presented above may be unintuitive. Then, hyperparameters could be elicited by consideration of (a) similarity between individual networks G^j, G^k, and (b) concordance of individual networks G^j with the prior network G^0. Specifically, we suggest the following two-step procedure: (a) exploit the fact that (for an uniform prior on G) we have $s_1 := p(i \notin G_p^j \Delta G_p^k) = 1 - 2h_\lambda + 2h_\lambda^2$, which facilitates selection of h_λ via the formula $h_\lambda = (1 + \sqrt{2s_1 - 1})/2$. (b) elicit h_η using the observation that $s_2 := p(i \notin G_p^j \Delta G_p^0) = 1 - h_\lambda - h_\eta + 2h_\lambda h_\eta$, so that $h_\eta = (s_2 + h_\eta - 1)/(2h_\lambda - 1)$. This two-step procedure uniquely determines a pair $(h_\eta, h_\lambda) \in [0.5, 1)^2$ and hence unique hyperparameters $(\eta, \lambda) \in [0, \infty)^2$. One drawback of this approach is that λ is selected under an assumption of a uniform prior on G; that is, $\eta = 0$. The quality of this procedure will therefore depend on the actual informativeness η of the prior network G^0 on G selected in step (b). This approach to hyperparameter selection has an analogous interpretation using expected total SHD.

The above heuristics may be useful in setting hyperparameters in practice. However, these heuristics are certainly no panacea and should be accompanied by checks of sensitivity to hyperparameters, as we report below.

3 Joint network inference for time-course data

The JNI model and network priors, as described above, are general. To apply the JNI framework in a particular context requires an appropriate likelihood at the individual level, that is, to specify the distribution $p(\boldsymbol{y}^j | G^j, \boldsymbol{\theta}^j)$ of data $\mathbf{y}^j$ conditional on graph G^j and parameters $\boldsymbol{\theta}^j$. In this Section we focus on time-course data, using DBNs to provide the likelihood.

3.1 Dynamic Bayesian network formulation

A DBN is a graphical model based on a DAG whose vertices have explicit time indices; see Murphy (2002) for details. Here, following Hill *et al.* (2012) and others, we use stationary DBNs and permit only edges forwards in time. Background and assumptions for DBNs are described in Appendix A. Further assuming a modular network prior, structural inference for DBNs can be carried out efficiently, as described in detail in Hill *et al.* (2012). A novel contribution of this paper is to extend these results to allow for efficient and exact *joint* estimation. In order to simplify notation, we define a data-dependent functional

$$\mathfrak{P}(\boldsymbol{X}) = p(\boldsymbol{X}(1)) \prod_{i=2}^{m} p(\boldsymbol{X}(i) | \boldsymbol{y}(i-1)) \tag{6}$$

which implicitly conditions upon observed history. Let $y_p^j(t)$ denote the observed value of variable p in individual j at time t. The above notation allows us to conveniently summarize the product

$$p(y_p^j(1) | G_p^j) p(y_p^j(2) | \boldsymbol{y}(1), G_p^j) \dots p(y_p^j(m) | \boldsymbol{y}(m-1), G_p^j). \tag{7}$$

as $\mathfrak{P}(\boldsymbol{y}_p^j | G_p^j)$. Thus, we have that, for DBNs, the full likelihood also satisfies modularity:

$$p(\boldsymbol{y} | G^1, \dots, G^J) \quad = \quad \prod_{j \in \mathcal{J}} \prod_{p \in \mathcal{P}} \mathfrak{P}(\boldsymbol{y}_p^j | G_p^j) \tag{8}$$

In other words, the parent sets G_p^j ($p \in \mathcal{P}$, $j \in \mathcal{J}$) are mutually orthogonal in the Fisher sense, so that inference for each may be performed separately.

For this paper, the local Bayesian score $\mathfrak{P}(\boldsymbol{y}_p^j | G_p^j)$ corresponds to the marginal likelihood for a linear autoregressive formulation described in Appendix B. We consider an extension to facilitate the analysis of datasets which contain interventions; this is described in Appendix C. For this choice of model it is possible to construct a fully conjugate set of priors, delivering a closed form expression for the local score, contained in Appendix D.

3.2 Computationally efficient joint estimation

Previous studies have used MCMC to generate samples from the posterior distribution over networks (Penfold *et al.*, 2012; Werhli and Husmeier, 2008). However, ensuring mixing has proven to be extremely challenging for joint estimation, with both studies reporting extremely slow convergence. Advances in MCMC and parallel computing may in the future ameliorate these issues (Lee *et al.*, 2010), but at present it remains the case that fast, interactive joint estimation is currently challenging or prohibitively demanding using MCMC. We therefore propose an exact approach, using an in-degree restriction coupled with prior modularity and a sum-product-type algorithm, to facilitate efficient estimation. For example, the DREAM4 problem ($P = 10$ variables, $J = 5$ individuals) considered by Penfold *et al.* (2012) was reported to require "several hours per node" for MCMC convergence; our approach solves the entire problem in ≈ 3 seconds. Our approach therefore complements MCMC-based inference, allowing fast, interactive investigation in moderate-dimensional settings.

Specifically, we use exact model averaging to marginalize over graphs and report posterior marginal inclusion probabilities. We begin by computing and caching the marginal likelihoods $\mathfrak{P}(\boldsymbol{y}_p^j|G_p^j)$ for all parent sets $G_p^j \in \mathcal{G}_p$, all variables $p \in \mathcal{P}$ and all individuals $j \in \mathcal{J}$; these could be obtained using essentially any suitable likelihood. The posterior marginal probability for an edge (i, p) belonging to the latent network G is computed as

$$\begin{aligned}
p(i \in G_p|\boldsymbol{y}, G^0) &= \sum_{G_p \in \mathcal{G}_p} \mathbf{1}_{i \in G_p} p(G_p|\boldsymbol{y}_p, G_p^0) \quad (9)\\
&= \sum_{G_p \in \mathcal{G}_p} \mathbf{1}_{i \in G_p} \sum_{G_p^j \in \mathcal{G}_p : j \in \mathcal{J}} p(G_p^1, \ldots, G_p^J, G_p|\boldsymbol{y}_p, G_p^0) \quad (10)\\
&\propto \sum_{G_p \in \mathcal{G}_p} \mathbf{1}_{i \in G_p} \sum_{G_p^j \in \mathcal{G}_p : j \in \mathcal{J}} \mathfrak{P}(\boldsymbol{y}_p|G_p^1, \ldots, G_p^J, G_p, G_p^0) p(G_p^1, \ldots, G_p^J, G_p|G_p^0) \quad (11)\\
&= \sum_{G_p \in \mathcal{G}_p} \mathbf{1}_{i \in G_p} \sum_{G_p^j \in \mathcal{G}_p : j \in \mathcal{J}} p(G_p|G_p^0) \prod_{j \in \mathcal{J}} \mathfrak{P}(\boldsymbol{y}_p|G_p^j) p(G_p^j|G_p) \quad (12)\\
&= \sum_{G_p \in \mathcal{G}_p} \mathbf{1}_{i \in G_p} p(G_p|G_p^0) \sum_{G_p^j \in \mathcal{G}_p : j \in \mathcal{J}} \prod_{j \in \mathcal{J}} \mathfrak{P}(\boldsymbol{y}_p|G_p^j) p(G_p^j|G_p) \quad (13)\\
&= \sum_{G_p \in \mathcal{G}_p} \mathbf{1}_{i \in G_p} p(G_p|G_p^0) \prod_{j \in \mathcal{J}} \sum_{G_p^j \in \mathcal{G}_p} \mathfrak{P}(\boldsymbol{y}_p|G_p^j) p(G_p^j|G_p) \quad (14)
\end{aligned}$$

where Eqn. 14 uses the sum-product lemma (Kschischang *et al.*, 2001) to interchange operators (see Appendix E). This final step has important consequences for algorithmic complexity (see Section 3.3). Note that, whilst this derivation can made without the explicit marginalization of Eqn. 10, the approach is quite general and may be used analogously to facilitate estimation of individual networks G^j:

$$\begin{aligned}
&p(i \in G_p^j|\boldsymbol{y}, G^0)\\
&= \sum_{G_p^j \in \mathcal{G}_p} \mathbf{1}_{i \in G_p^j} p(G_p^j|\boldsymbol{y}_p, G_p^0) \quad (15)\\
&= \sum_{G_p^j \in \mathcal{G}_p} \mathbf{1}_{i \in G_p^j} \sum_{G_p \in \mathcal{G}_p} \sum_{G_p^k \in \mathcal{G}_p : k \in \mathcal{J} \setminus \{j\}} p(G_p^1, \ldots, G_p^J, G_p|\boldsymbol{y}_p, G_p^0) \quad (16)\\
&\propto \sum_{G_p^j \in \mathcal{G}_p} \mathbf{1}_{i \in G_p^j} \sum_{G_p \in \mathcal{G}_p} \sum_{G_p^k \in \mathcal{G}_p : k \in \mathcal{J} \setminus \{j\}} p(\boldsymbol{y}_p|G_p^1, \ldots, G_p^J, G_p, G_p^0) p(G_p^1, \ldots, G_p^J, G_p|G_p^0) \quad (17)\\
&= \sum_{G_p^j \in \mathcal{G}_p} \mathbf{1}_{i \in G_p^j} \sum_{G_p \in \mathcal{G}_p} \sum_{G_p^k \in \mathcal{G}_p : k \in \mathcal{J} \setminus \{j\}} p(G_p|G_p^0) \prod_{l \in \mathcal{J}} \mathfrak{P}(\boldsymbol{y}_p^l|G_p^l) p(G_p^l|G_p) \quad (18)\\
&= \sum_{G_p^j \in \mathcal{G}_p} \mathbf{1}_{i \in G_p^j} \sum_{G_p \in \mathcal{G}_p} p(G_p|G_p^0) \mathfrak{P}(\boldsymbol{y}_p^j|G_p^j) p(G_p^j|G_p) \sum_{G_p^k \in \mathcal{G}_p : k \in \mathcal{J} \setminus \{j\}} \prod_{l \in \mathcal{J} \setminus \{j\}} \mathfrak{P}(\boldsymbol{y}_p^l|G_p^l) p(G_p^l|G_p) \quad (19)\\
&= \sum_{G_p^j \in \mathcal{G}_p} \mathbf{1}_{i \in G_p^j} \sum_{G_p \in \mathcal{G}_p} p(G_p|G_p^0) \mathfrak{P}(\boldsymbol{y}_p^j|G_p^j) p(G_p^j|G_p) \prod_{k \in \mathcal{J} \setminus \{j\}} \sum_{G_p^k \in \mathcal{G}_p} \mathfrak{P}(\boldsymbol{y}_p^l|G_p^l) p(G_p^l|G_p) \quad (20)
\end{aligned}$$

where again the sum-product lemma justifies the exchange of operators.

3.3 Computational complexity

Following Hill *et al.* (2012) we reduced the space of parent sets $\mathcal{G}_p$ using an in-degree restriction of the form $|G_p^j| \leq c$ for all $G_p^j \in \mathcal{G}_p$, $p \in \mathcal{P}$, $j \in \mathcal{J}$. Thus the cardinality of the space of parent sets $M = |\mathcal{G}_p| = \mathcal{O}(P^c)$ is polynomial in P, where it was previously exponential. This reduces summation over an exponential number of terms to a more manageable sum over polynomially many terms. Moreover, in the protein signaling example to follow, bounded in-degree is a reasonable biological assumption. Sensitivity to choice of c is discussed in Section 4.1.

Caching of selected probabilities is used to avoid redundant recalculation. Pseudocode is provided in Appendix F, which consists of three phases of computation. Storage costs are dominated by Phases I and II, which each requiring the caching of $\mathcal{O}(PJM)$ real numbers. Phase II dominates computational effort, with total (serial) algorithmic complexity $\mathcal{O}(PJ^2M^2)$. However, within-phase computation is "embarrassingly parallel" in the sense that all calculations are independent (indicated by square parentheses notation in the pseudocode). Thus an ideal implementation requires $\mathcal{O}(3)$ computational time. We provide a MATLAB implementation in Supplement B.

4 Results

We tested our joint estimation procedure on simulated time-course data. We compare our approach to the special cases of (i) inferring each network separately (INI); (ii) allowing parameters (but not networks) to change between individuals (ANI); (iii) the naive approach of aggregating all data (monolithic) and (iv) simple temporal correlations (absolute Pearson coefficient). For a fair comparison, all methods, with the exception of (iv), were implemented so as to take account of the interventional nature of the data. We note that it is not possible to directly compare our results with Danaher *et al.* (2012); Guo *et al.* (2011); Yang *et al.* (2012) since these methods do not apply to time-course data. The method of Penfold *et al.* (2012) applies to time-course data, but the computational demands of the approach precluded application in this setting. Specifically, in the simulated data example we report below, over 3000 rounds of inference were performed in total, on problems larger than DREAM4 ($P = 10$, $J = 5$). Using the approach of Penfold *et al.* (2012), these experiments would have required more than 10 years computational time; in contrast our approach required less than 24 hours serial computation on a standard laptop.

4.1 Performance metrics

The proposed methodology addresses three questions, some or all of which may be of scientific interest depending on application; (i) estimation of the latent network G, (ii) estimation of individual networks $G^1, \dots, G^J$, and (iii) estimation of differences between individual networks. We quantify performance for tasks (i) and (ii) using the area under the receiver operating characteristic (ROC) curve (AUR). This metric, equivalent to the probability that a randomly chosen true edge is preferred by the inference scheme to a randomly chosen false edge, summarizes, across a range of thresholds, the ability to select edges in the data-generating network. AUR may be computed relative to the true latent network G, or relative to the true individual networks G^j, quantifying performance on tasks (i) and (ii) respectively. Both sets of results are presented below, in the latter case averaging AUR over all individual networks. For (iii), in order to assess ability to estimate individual heterogeneity, we computed AUR scores based on the statistics $F_{ip}^j = |p(i \in G_p^j|\boldsymbol{y}, G^0) - p(i \in G_p|\boldsymbol{y}, G^0)|$ which should be close to one if $i \in G_p^j \Delta G_p$, otherwise F_{ip}^j should be close to zero.

It is easy to show that inference for the latent network, under only the prior, attains mean AUR equal to h_η. Similarly, prior inference for the individual networks attains mean AUR equal to $1 - h_\eta - h_\lambda + 2h_\eta h_\lambda$. This provides a baseline for the proposed methodology at tasks (i) and (ii) and allows performance to be decomposed into AUR due to prior knowledge and AUR contributed through inference. Using a systematic variation of data-generating parameters, we defined 15 distinct data generating regimes. For all 15 regimes we considered 50 independent datasets; standard errors accompany average AUR scores. Results presented below use a computationally favorable in-degree restriction $c = 3$. Note that when the maximum in-degree of any of the true networks exceeds the computational restriction c, estimator consistency will not be guaranteed. In order to check robustness to c, a subset of experiments were repeated using $c = 4$, with close agreement observed (SFig. 4).

4.2 Data generation

A latent network G on P vertices was drawn from the Erdös distribution with edge density ρ/P. In order to simulate heterogeneity, the individual networks G^j were obtained from G by maintaining the status (present/absent) of each edge independently with probability h_λ. A parameter β^j_{ip} for each parent $i \in G^j_p$ was independently drawn from the mixture normal distribution $0.5\mathcal{N}(-1, 0.1^2) + 0.5\mathcal{N}(1, 0.1^2)$ (the mixture distribution ensures that parameters are not vanishingly small, so that the structural inference problem is well-defined). Collecting together parameters produces matrices $\boldsymbol{\beta}^j$, corresponding to networks G^j via $i \in G^j_p$ if and only if $\beta^j_{ip} \neq 0$. We also generate, for each individual j, intercept parameters $\boldsymbol{\alpha}^j \sim N(\mathbf{0}_P, \boldsymbol{I}_{P\times P})$ representing baseline expression levels. Initial conditions were sampled as $\boldsymbol{y}^j(1) \sim N(\mathbf{0}_P, \boldsymbol{I}_{P\times P})$. Data were then generated from the autoregressive model $\boldsymbol{y}^j(t) = \boldsymbol{\alpha}^j + \boldsymbol{y}^j(t-1)\boldsymbol{\beta}^j + \boldsymbol{\epsilon}_t$, where $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}_P, \sigma^2 \boldsymbol{I}_{P\times P})$ are independent for $t = 2, \ldots, n$. In this way E such time courses were obtained; that is, from E distinct initial conditions, so the total number of data for individual j is $n_j = En$. In order to avoid issues of blow-up and to generate plausible datasets, the matrices $\boldsymbol{\beta}^j$ were normalized by their spectral radii prior to data generation.

In order to investigate the effect of using a prior network G^0, we do not simply want to set G^0 equal to the latent network G, since in practice this network is unknown. We therefore generated a prior network G^0 by correctly specifying each potential edge as either present or absent with probability h_η. In this way we mimic partial prior knowledge of the networks under study.

4.3 Alternative data generating mechanisms

We augmented the above data-generating scheme to mimic interventional experiments. In this case, for each time course, a randomly chosen variable is marked as the target of an interventional treatment. Data are then generated according to the augmented likelihood described in Appendix C (fixed effects were taken to be zero). Furthermore, in order to investigate the impact of model misspecification, we also considered time series data generated from a computational model of protein signaling, based on nonlinear ODEs (Xu *et al.*, 2010). In order to extend this model, which is for a single cell type, to simulate a heterogeneous population, we randomly selected three protein species per individual and deleted their outgoing edges in the data-generating network (see Supplement A).

4.4 Latent network

Firstly we investigated ability to recover the latent network G. Initially all estimators are assigned approximately optimal hyperparameter values $\eta = 1, \lambda = 4$ (for Xu *et al.* (2010), $\lambda = 3$) based on the heuristic of Eqn. 4; prior misspecification is investigated later in Section 4.7. We found little difference in the ability of JNI and ANI to recover the latent network structure across a wide range of regimes (STable 1). Since ANI enjoys favorable computational complexity, this estimator may be preferred for this task in practice. However, both approaches clearly outperformed monolithic inference, which was no better than inference based on the prior alone, demonstrating the importance of accounting for variation in parameter values. Correlations barely outperformed random sampling.

In practice, one could also estimate G using independent network inference (INI), via the *ad hoc* estimator $p(i \in G_p|\boldsymbol{y}, G^0) \approx \frac{1}{J}\sum_{j\in\mathcal{J}} p(i \in G^j_p|\boldsymbol{y}^j, G^0)$ which performs an unweighted average of J independent network inferences. However we found that INI offered no advantage over JNI and ANI, performing worse than both in

14 out of 15 regimes. We obtained qualitatively similar results for both alternative data-generating schemes (STables 3,6).

4.5 Individual networks

Secondly we investigated the ability to recover individual networks G^j. At this task, JNI outperformed INI in all 15 regimes (Table 1). This demonstrates a substantial increase in statistical power resulting from the hierarchical Bayesian approach. JNI also outperformed monolithic estimation and inference using temporal correlations in all 15 regimes, with the latter demonstrating substantial bias.

One may try to improve upon INI by firstly estimating the wild type network G, and then taking this estimate as a prior network G^0 within a second round of INI. Informed by Section 4.4, we consider the approach whereby G is first estimated using ANI, referring to this two-step procedure as "empirical network inference" (ENI). We found that the performance of ENI consistently matched that of JNI over a wide range of regimes. Since ENI avoids all joint computation, this may provide a practical estimator of individual networks in higher dimensional settings. Similar results were observed using the alternative data-generating schemes, although JNI slightly outperformed ENI on the Xu *et al.* (2010) datasets (STables 4,7).

4.6 Feature detection

Thirdly, we assessed ability to pinpoint sources of variation within the population. Interest is often directed toward individual-specific heterogeneity, or *features*. Informally, writing $G^j = G + \delta^j$, features correspond to δ^j. JNI regularizes between individuals; it therefore ought to eliminate spurious differences, leaving only features which are strongly supported by data. Equivalently, since JNI offers improved estimation of the latent network G, the features $\delta^j = G^j - G$ ought also to be better estimated.

Feature detection may also be performed using INI or ENI, comparing an latent network estimator (see *ad hoc* estimator in Section 4.4) with individual networks. The performance of JNI was compared to the performance of INI and ENI (STable 2). We found that, whilst feature detection is much more challenging that previous tasks, JNI mostly outperformed both INI and ENI, with exceptions occurring whenever the underlying dataset was highly informative (in which case INI was often superior). This suggests that coherence of the JNI analysis aids in suppressing spurious features in the small sample setting. Alternative data-generating schemes produced qualitatively similar results, although JNI outperformed ENI on the Xu *et al.* (2010) datasets (STables 5,8).

4.7 Robustness to hyperparameter misspecification

For the above investigation we used Eqn. 4 to elicit hyperparameters λ, η. This was possible because the data-generating parameters h_λ, h_η were known by design; however in general this will not be the case. It is therefore important that estimator performance does not deteriorate heavily when alternative hyperparameter values are employed. By fixing (h_λ, h_η) in the data generating process, we are able to investigate the robustness of JNI estimator to hyperparameter misspecification. In particular, when finite values are ascribed to data-generating parameters (h_η, h_λ), ANI and INI may be interpreted as inference using misspecified prior distributions (see Section 2.3).

SFig. 3 displays how performance of the JNI estimator for latent networks depends on the choice of hyperparameters $\lambda, \eta \in [0, \infty)$. We notice that AUR remains close to that obtained for optimal λ, η over a fairly large interval, so that performance is not exquisitely dependent on prior elicitation.

Data Generating Regime								Estimator					
J	n	E	P	σ	ρ	h_η	h_λ	JNI	ANI	INI	Monolithic	Correl.	Prior
10	5	1	10	0.2	1	0.73	0.98	0.88 ± 0.0088	0.73 ± 0.011	0.87 ± 0.01	0.71 ± 0.012	0.55 ± 0.013	0.72
5	5	1	10	0.2	1	0.73	0.98	0.86 ± 0.0083	0.74 ± 0.01	0.85 ± 0.0092	0.75 ± 0.01	0.55 ± 0.015	0.72
20	5	1	10	0.2	1	0.73	0.98	0.88 ± 0.0057	0.74 ± 0.0098	0.88 ± 0.0074	0.68 ± 0.0089	0.59 ± 0.015	0.72
10	10	1	10	0.2	1	0.73	0.98	0.94 ± 0.0051	0.86 ± 0.0075	0.95 ± 0.0051	0.63 ± 0.012	0.56 ± 0.015	0.72
10	5	5	10	0.2	1	0.73	0.98	0.97 ± 0.0035	0.94 ± 0.0052	0.98 ± 0.0041	0.7 ± 0.011	0.6 ± 0.014	0.72
10	5	1	20	0.2	1	0.73	0.98	0.86 ± 0.0046	0.78 ± 0.0075	0.86 ± 0.0057	0.67 ± 0.0072	0.54 ± 0.0078	0.72
10	5	1	10	0.1	1	0.73	0.98	0.88 ± 0.009	0.75 ± 0.0094	0.88 ± 0.011	0.71 ± 0.012	0.53 ± 0.017	0.72
10	5	1	10	1	1	0.73	0.98	0.81 ± 0.0089	0.7 ± 0.0093	0.79 ± 0.013	0.72 ± 0.0084	0.51 ± 0.013	0.72
10	5	1	10	0.2	0.5	0.73	0.98	0.84 ± 0.012	0.67 ± 0.017	0.84 ± 0.013	0.69 ± 0.017	0.56 ± 0.016	0.72
10	5	1	10	0.2	2	0.73	0.98	0.88 ± 0.0068	0.73 ± 0.0087	0.84 ± 0.0089	0.7 ± 0.0099	0.54 ± 0.01	0.72
10	5	1	10	0.2	1	0.62	0.98	0.86 ± 0.0087	0.63 ± 0.012	0.86 ± 0.009	0.64 ± 0.013	0.53 ± 0.015	0.62
10	5	1	10	0.2	1	0.88	0.98	0.9 ± 0.0052	0.88 ± 0.0066	0.89 ± 0.0088	0.79 ± 0.0089	0.55 ± 0.011	0.87
10	5	1	10	0.2	1	0.73	0.73	0.57 ± 0.0041	0.56 ± 0.0043	0.56 ± 0.0044	0.54 ± 0.0037	0.52 ± 0.0036	0.61
10	5	1	10	0.2	1	0.73	1	0.9 ± 0.014	0.75 ± 0.019	0.9 ± 0.012	0.73 ± 0.014	0.56 ± 0.016	0.73
10	5	1	10	0.2	1	0.73	0.98	0.88 ± 0.0084	0.74 ± 0.012	0.89 ± 0.0095	0.71 ± 0.011	0.55 ± 0.013	0.72

Table 1: Assessment of estimators for inference of individual networks G^j; autoregressive dataset with interventions. [Values shown are average AUR ± standard error, over 50 realizations. Green/red is used to indicate the highest/lowest scoring estimators. J = number of individuals, n = number of time points per time course, E = number of time courses, P = number of variables, σ = noise magnitude, (h_η, h_λ) = data generating hyperparameters. "JNI" = joint network inference, "ANI" = aggregate data but control for parameter confounding, "INI" = average J independent network inferences, "Monolithic" = aggregate data without controlling for parameter confounding, "Correl." = estimation using the absolute Pearson temporal correlation coefficient, "Prior" = estimation using only the prior network G^0.]

4.8 Robustness to outliers and batch effects

The biological datasets which motivate this study often contain outliers. At the same time, experimental design may lead to platform-specific batch effects. In order to probe estimator robustness, we generated data as previously described, with the addition of outliers and certain batch effects. Specifically, Gaussian noise from the contamination model $0.95\mathcal{N}(0, 0.1^2) + 0.05\mathcal{N}(0, 10^2)$ was added to all data prior to inference. At the same time, one individual's data were replaced entirely by Gaussian white noise to simulate a batch effect that could arise if preparation of a specific biological sample was incorrect. The relative decrease in performance at feature detection is reported in SFig. 5. We found that JNI remained the optimal estimator for all three estimation problems, in spite of these heavy violations to the modeling assumptions. However, the actual decrease in performance was more pronounced for JNI than for INI, suggesting that decoupled estimation (INI) may confer robustness to batch effects which affect single individuals.

5 Discussion

There are three distinct, though related, structure learning problems which may be addressed in the context of an heterogeneous population of individuals:

1. Recovering a shared or "wild type" network from the heterogeneous data.

2. Recovering networks for specific individuals.

3. Pinpointing network variation within the population.

Each problem may be of independent scientific interest, and the joint approaches investigated here address all three problems simultaneously within a coherent framework. We considered simulated data, with and without model misspecification. For all three problems we demonstrated that a joint analysis performs at least as well as independent or aggregate analyses.

Our analysis, based on exact Bayesian model averaging, was massively faster then the sampling-based schemes of Penfold *et al.* (2012); Werhli and Husmeier (2008). Moreover, our estimators are deterministic, so that difficulties pertaining to MCMC convergence were avoided. Indeed, attaining convergence on joint models of this kind appears to be challenging (Werhli and Husmeier, 2008). The proposed methodology is scalable, with an embarrassingly parallel algorithm provided in Section 3.3. Furthermore, we described approximations to a joint analysis which enjoy further reduced computational complexity whilst providing almost equal estimator performance across a wide range of data-generating regimes.

Whilst we considered the simplest form of regularization, based on prior modularity, there is potential to integrate richer knowledge into inference. One possibility would be hierarchical regularization that allows entire pathways to be either active or inactive. However, in this setting it would be important to revisit hyperparameter elicitation; the procedures which we have described are specific to SHD priors. In particular we restricted the joint model to have equal inverse temperatures $\lambda^1 = \cdots = \lambda^J := \lambda$. Relaxing this assumption may improve robustness to batch effects which target single individuals, since then weak informativeness ($\lambda^j \approx 0$) may be learned from data. It would also be interesting to distinguish between $G \setminus G^j$ ("loss of function") and $G^j \setminus G$ ("gain of function") features. In this work we did not explore information sharing through parameter values $\boldsymbol{\theta}^j$, yet this may yield more powerful estimators of network structure in settings where individuals' parameters $\boldsymbol{\theta}^j, \boldsymbol{\theta}^k$ are not independent.

The JNI model could be formulated as a penalized (log-)likelihood

$$\log(p(\boldsymbol{y}|G^1, \ldots, G^J)) - \sum_{j \in \mathcal{J}} \lambda^j d^j(G^j, G; A^j) - \eta d(G, G^0). \tag{21}$$

The frequentist approaches described by Danaher *et al.* (2012); Guo *et al.* (2011); Yang *et al.* (2012) enjoy favorable computational complexity (esp. Danaher *et al.* (2012) who provide an example with $P = 22,283$ variables and $J = 187$ individuals). However, in small to moderate dimensional settings, the Bayesian methods proposed here are complementary in several respects: (i) Bayesian approaches provide a confidence measure for inferred topology, dealing with non-identifiable and multi-modal problems; (ii) no convexity assumptions are required on the form of the penalty functions d, d^j in the Bayesian setting, which may assist with integration of ancillary information; (iii) the above penalized likelihood methods do not apply directly to time course data (but could be extended to do so).

These experiments employed a promising formulation of likelihood under intervention due to Spencer and Mukherjee (2012). There are a number of interesting extensions which may be considered in future work: (i) In high dimensions, Bayesian variable selection requires multiplicity correction in order to avoid degeneracy (Scott and Berger, 2010). Such correction is required to control the false discovery rate and is independent to the penalty on model complexity provided by the marginal likelihood. In this moderate-dimensional work, in order to simplify the presentation, we did not employ a multiplicity correction; this should be an avenue for future development. (ii) Inference was based upon a local score borrowed from Bayesian linear regression. We chose to employ the g-prior due to Zellner (1986), where following George and Foster (2000) we used (conditional) empirical Bayes to select the g hyperparameter. Others have suggested setting $g = n$ (unit information prior; Smith *et al.*, 2001), whilst Deltell *et al.* (2012) and Liang *et al.* (2008) propose prior distributions over g with attractive theoretical properties. Our empirical investigation suggested that the choice of hyperparameter elicitation is influential, but a thorough comparison of linear model specifications is beyond the scope of this paper. (iii) As discussed in Oates and Mukherjee (2012), linear autoregressive formulations may be inadequate in realistic settings; in particular, samples which are obtained unevenly in time can be problematic. Recent advances which incorporate mechanistic detail into the likelihood may prove advantageous (Oates *et al.*, 2012). Since the JNI approach decouples the marginal likelihood and model averaging computations, it may be applied directly to the output of more sophisticated models. (iv) In the case of linear models, Barbieri and Berger (2004) showed that the median probability model (i.e. model averaging) provides superior predictive performance over the maximum *a posteori* (MAP) model. However we are unaware of an analogous result for causal inference in the Bayesian setting.

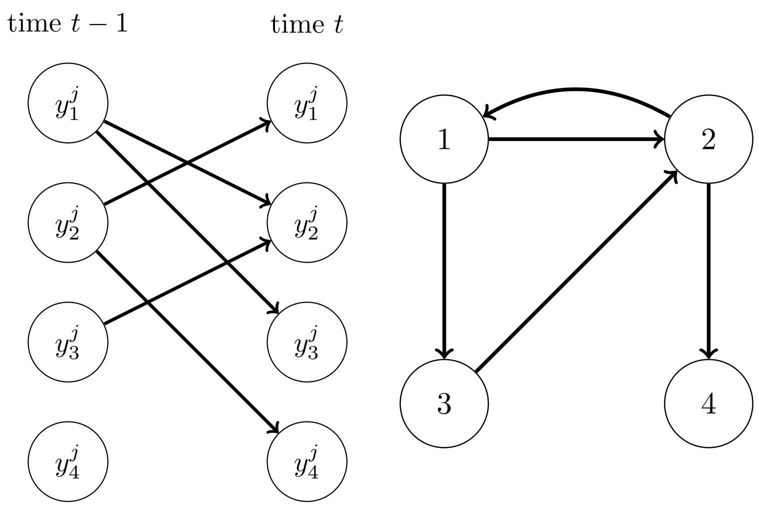
Techniques for modeling heterogeneous data are clearly widely applicable. The methodology presented here may be applicable in other disciplines. For example, our approach is suited to meta-analyses of network analyses (Weile *et al.*, 2012), integration of multiple data sources (Kato *et al.*, 2005; Wei and Pan, 2012; Werhli and Husmeier, 2008) or data arising from context dependent networks (Baumbach *et al.*, 2009). The ideas discussed here share many connections with time-heterogeneous DBNs which, for brevity, we did not discuss in this paper (Dondelinger *et al.*, 2010, 2012; Grzegorczyk and Husmeier, 2011; Song *et al.*, 2009).

Acknowledgements

We would like to thank J.D. Aston, F. Dondelinger, C.A. Penfold, S.E.F. Spencer and S.M. Hill for helpful discussion and comments. Financial support was provided by NCI U54 CA112970, UK EPSRC EP/E501311/1 and the Cancer Systems Center grant from the Netherlands Organisation for Scientific Research.

Appendix A: Dynamic Bayesian networks

DBNs have emerged as popular tools for the analysis of multivariate time course data due to (i) the fact that no acyclicity assumption is required on the (static) network, and (ii) computational tractability resulting from a factorization of the likelihood function over variables $p \in \mathcal{P}$ (Hill *et al.*, 2012). For the DBNs used here, an edge (p, q) from $p \in \mathcal{P}$ to $q \in \mathcal{P}$ in $G^j \in \mathcal{G}$ means that $y_q^j(t)$, the observed value of variable q in individual j at time t, depends directly upon $y_p^j(t-1)$, the observed value of p in individual j at time $t-1$ (Fig. 3(a); note that t indexes sample index, rather than actual sampling time). Let $\boldsymbol{y}^j$ denote a vector containing all observations for individual j. Then $\boldsymbol{y}^j(t)$ is conditionally independent of $\{\boldsymbol{y}^j(t-\tau) : \tau \geq 2\}$ given $\boldsymbol{y}^j(t-1)$ and G^j (first-order Markov assumption). These conditional independence relations are conveniently summarized as a (static) network G^j with exactly P vertices (Fig. 3(b)); note that this latter network need not be acyclic.



(a) G^j; DBN representation. (b) G^j; "static" representation.

Figure 3: Dynamic Bayesian networks (DBNs). (a) An "unrolled" dynamic Bayesian network (DBN) showing each variable at successive time points. (b) The corresponding "static" representation of DBN (a) with exactly one vertex for each variable.

Appendix B: Local likelihood for time course data

We follow Aliferis *et al.* (2010); Hill *et al.* (2012); Penfold *et al.* (2012) in formulating inference in DBNs as a regression problem. We entertain models for the response $y_p^j(t)$ as predicted by covariates $\boldsymbol{y}^j(t-1)$. In many cases multiple time series will be available. In this case the vector $\boldsymbol{y}_p^j$ contains the concatenated time series. The DBN formulation gives rise to the following linear regression likelihood

$$\boldsymbol{y}_p^j = \boldsymbol{X}_0\boldsymbol{\alpha} + \boldsymbol{X}^j_{G_p^j}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{22}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}_{n\times 1}, \sigma^2\boldsymbol{I}_{n\times n})$. The matrix $\boldsymbol{X}_0 = [\mathbf{1}_{\{t=1\}}\ \mathbf{1}_{\{t>1\}}]_{n\times 2}$ contains a term for the initial time point in each experiment. The elements of $\boldsymbol{X}^j_{G_p^j}$ corresponding to initial observations $y_p^j(1)$ are simply set to zero. Parameters $\boldsymbol{\theta}_p^j = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma\}$ are specific to model G_p^j, variable p and cell line j. In the simplest case the model-specific component $\boldsymbol{X}^j_{G_p^j}$ of the design matrix consists of the raw predictors $\boldsymbol{y}^j_{G_p^j}(t-1)$ where $\boldsymbol{y}_A^j$ denotes the elements of the vector $\boldsymbol{y}^j(t-1)$ belonging to the set A, though more complex basis functions may be used.

Appendix C: Modeling interventions

Following Eaton and Murphy (2007); Spencer and Mukherjee (2012) we model interventional data by modification to the DAG in line with a causal calculus (Pearl, 2009). We mention briefly some of the key ideas and refer the interested reader to the references for full details. A "perfect intervention" corresponds to 100% removal of the target's activity with 100% specificity. In the context of protein phosphorylation, kinases may be intervened upon using agents such as monoclonal antibodies, small molecule inhibitors or even si-RNA (Lu *et al.*, 2011). We make the simplifying assumptions that these interventions are perfect and use the "perfect out fixed effects" (POFE) approach recommended by Spencer and Mukherjee (2012). We refer the reader to Spencer and Mukherjee (2012) for an extended discussion of POFE. This changes the DAG structure to model the intervention and also estimates a fixed effect parameter to model the change under intervention in the log-transformed data.

Appendix D: Exact marginal likelihood

We employed a Jeffreys prior $p(\boldsymbol{\alpha}, \sigma | G_p^j, \phi^j) \propto 1/\sigma$ for $\sigma > 0$ over the common parameters. Prior to inference, the non-interventional components of the design matrix were orthogonalized using the transformation $(\boldsymbol{X}_{G_p^j}^j)_{ik} \mapsto \sum_l (\boldsymbol{I}_n - \boldsymbol{P}_0)_{il} (\boldsymbol{X}_{G_p^j}^j)_{lk}$, where $\boldsymbol{P}_0 = \boldsymbol{X}_0 (\boldsymbol{X}_0^T \boldsymbol{X}_0)^{-1} \boldsymbol{X}_0^T$ (Deltell *et al.*, 2012). We then assumed a g-prior for regression coefficients (Zellner, 1986), given by $\boldsymbol{\beta} | \boldsymbol{\alpha}, \sigma, G_p^j, \phi^j \sim N(\boldsymbol{0}_{b\times 1}, \phi^j \sigma^2 (\boldsymbol{X}_{G_p^j}^T \boldsymbol{X}_{G_p^j})^{-1})$ where $b = \dim(\boldsymbol{\beta})$. Using these priors for the DBNs with intervention as outlined above, the marginal likelihood can be obtained in closed-form:

$$\mathfrak{P}(\boldsymbol{y}_p^j | G_p^j, \phi^j) \propto \frac{1}{(\phi^j + 1)^{b/2}} \left(\boldsymbol{y}_p^{jT} \left(\boldsymbol{I}_{n\times n} - \boldsymbol{P}_0 - \frac{\phi^j}{\phi^j + 1} \boldsymbol{P}_{G_p^j} \right) \boldsymbol{y}_p^j \right)^{-\frac{n-a}{2}} \tag{23}$$

where $\boldsymbol{P}_{G_p^j} = \boldsymbol{X}_{G_p^j} (\boldsymbol{X}_{G_p^j}^T \boldsymbol{X}_{G_p^j})^{-1} \boldsymbol{X}_{G_p^j}^T$, $a = \dim(\boldsymbol{\alpha})$ and $b = \dim(\boldsymbol{\beta})$. Empirical investigations have previously demonstrated good results for network inference based on the above marginal likelihood (Hill *et al.*, 2012; Spencer and Mukherjee, 2012). Following George and Foster (2000) we used the (conditional) empirical Bayes approach to determine the ϕ^j hyperparameter (details in Supplement A).

Appendix E: The sum-product lemma

The "sum-product" lemma, which forms the basis for several exact inference procedures in graphical models, can be expressed in its most basic form as follows: For a finite set of functionals $f_i : \mathcal{X}_i \to \mathbb{R}$ on finite domains ϕ_i indexed by $1 \leq i \leq I$ we have

$$\sum_{x_1 \in \mathcal{X}_1, \ldots, x_I \in \mathcal{X}_I} \prod_{i=1}^{I} f_i(x_i) = \prod_{i=1}^{I} \sum_{x_i \in \mathcal{X}_i} f_i(x_i). \tag{24}$$

The proof is straight forward (induction on I) and can be found in e.g. Kschischang *et al.* (2001). The sum-product lemma is typically used to reduce algorithmic complexity, replacing the $\mathcal{O}(|\mathcal{X}_1| \times \cdots \times |\mathcal{X}_I| \times I)$ expression on the left hand side by the $\mathcal{O}(|\mathcal{X}_1| + \cdots + |\mathcal{X}_I|)$ expression on the right hand side.

Appendix F: Joint network inference - pseudocode

This appendix contains pseudocode for exact joint model averaging. [Computational complexity of calculating marginal likelihoods $\mathfrak{P}(\boldsymbol{y}_p^j | G_p^j)$ will scale with sample size n; scaling exponents shown here assume $\mathcal{O}(n) = \mathcal{O}(1)$.] Below we provide pseudocode for computation of posterior marginal inclusion probabilities for edges in individual networks G^j:

```
for all $p \in \mathcal{P}$ do
    Phase I:
      Compute and cache $[\forall p \in \mathcal{P}]$ $[\forall j \in \mathcal{J}]$ $[\forall G_p \in \mathcal{G}_p]$
      $\mathfrak{P}(\boldsymbol{y}_p^j | G_p) = \sum_{G_p^j \in \mathcal{G}_p} \mathfrak{P}(\boldsymbol{y}_p^j | G_p^j) p(G_p^j | G_p)$ $[\mathcal{O}(M)]$
    Phase II:
      Compute and cache $[\forall p \in \mathcal{P}]$ $[\forall j \in \mathcal{J}]$ $[\forall G_p^j \in \mathcal{G}_p]$
      $p(G_p^j | \boldsymbol{y}_p, G_p^0) \propto \sum_{G_p \in \mathcal{G}_p} p(G_p | G_p^0) \mathfrak{P}(\boldsymbol{y}_p^j | G_p^j) p(G_p^j | G_p) \prod_{k \in \mathcal{J} \setminus \{j\}} \mathfrak{P}(\boldsymbol{y}_p^j | G_p)$ $[\mathcal{O}(MJ)]$
    Phase III:
      Compute and cache $[\forall p \in \mathcal{P}]$ $[\forall j \in \mathcal{J}]$ $[\forall i \in \mathcal{P}]$
      $p(i \in G_p^j | \boldsymbol{y}, G^0) = \sum_{G_p^j \in \mathcal{G}_p} \boldsymbol{1}_{i \in G_p^j} p(G_p^j | \boldsymbol{y}, G_p^0)$ $[\mathcal{O}(M)]$
end for
```

Pseudocode for inference of the latent network G proceeds analogously.

References

Aliferis, C.F. *et al.* (2010) Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification, Part I: Algorithms and Empirical Evaluation. *J. Mach. Learn. Res.* **11**:171-234.

Bachman, K.E. *et al.* (2004) The PIK3CA gene is mutated with high frequency in human breast cancers, *Cancer Biol. Ther.* **3**(8):772-775.

Barbieri, M.M., Berger, J.O. (2004) Optimal predictive model selection, *Ann. Stat.* **32**(3):870-897.

Barretina *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**:603-607.

Baumbach *et al.* (2009) Reliable transfer of gene regulatory networks between taxonomically related organisms. *BMC Bioinformatics* **3**:8.

Bender, C. *et al* (2010) Dynamic deterministic effects propagation networks: learning signalling pathways from longitudinal protein array data. *Bioinformatics*, **26**(ECCB 2010):i596-i602.

Cao, J. *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* **43**:956-963.

Claassen, T., Heskes, T. (2012) A Bayesian Approach to Constraint Based Causal Inference. *Proceedings of the 28th Conference on Uncertainty and Artificial Intelligence, Santa Catalina CA USA.*

Danaher, P., Wang, P., Witten, D.M. (2012) The joint graphical lasso for inverse covariance estimation across multiple classes, arXiv:1111.0324 [stat.ME].

Davies *et al.* (2002) Mutations of the BRAF gene in human cancer. *Nature* **417**(6892):949-54.

Deltell, A. *et al.* (2012) Criteria for Bayesian Model Choice with Application to Variable Selection. *Ann. Stat.*, to appear.

Dondelinger, F., Lebre, S., Husmeier, D. (2010) Heterogeneous continuous dynamic Bayesian networks with flexible structure and inter-time segment information sharing. *Proceedings of the 27th International Conference on Machine Learning*, 303-310.

Dondelinger, F., Lebre, S., Husmeier, D. (2012) Non-homogeneous dynamic Bayesian networks with Bayesian regularization for inferring gene regulatory networks with gradually time-varying structure. *Mach. Learn.* **90**(2):191-230.

Eaton, D., Murphy, K. (2007) Exact Bayesian structure learning from uncertain interventions. *Proceedings of the 11th Conference on Artificial Intelligence and Statistics (AISTATS-07)*, 107-114.

Forbes *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucl. Acids Res.* **39**(Suppl 1):D945-D950.

George, E.I., Foster, D.P. (2000) Calibration and empirical Bayes variable selection. *Biometrika* **87**(4):731-747.

Grzegorczyk, M., Husmeier, D. (2011) Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes. *Bioinformatics* **27**(5):693-699.

Guo, J., Levina, E., Michailidis, G., Zhu, J. (2011) Joint estimation of multiple graphical models. *Biometrika* **98**:115.

Hennessey, B.T. *et al.* (2010) A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Nonmicrodissected Human Breast Cancer. *Clin. Proteom.* **6**:129-151.

Hill, S.M. *et al.* (2012) Bayesian Inference of Signaling Network Topology in a Cancer Cell Line. *Bioinformatics* **28**(21):2804-2810.

Hill, S.M. and Mukherjee, S. (2013) Network-based clustering with mixtures of L1-penalized Gaussian graphical models: an empirical investigation, arXiv:1301.2194 [stat.ML].

Hoff, P.D. (2009) A hierarchical eigenmodel for pooled covariance estimation. *J. Roy. Stat. Soc. B* **71**(5):971-992.

Ideker, T., Krogan, N.J. (2012) Differential network biology. *Mol Syst Biol* **8**:565.

Imoto *et al.* (2003) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Proc. IEEE Computer Society Bioinformatics Conference (CSB'03)*, 104-113.

Imoto *et al.* (2006) Error tolerant model for incorporating biological knowledge with expression data in estimating gene networks. *Stat. Method.* **3**(1):116.

Kalisch, M., Bühlmann, P. (2007) Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm. *J. Mach. Learn. Res.* **8**:613-636.

Kato, T., Tsuda, K., Asai, K. (2005) Selective integration of multiple biological data for supervised network inference. *Bioinformatics* **21**(10):2488-2495.

Kschischang, F.R., Frey, B.J., Loeliger, H.-A. (2001) Factor Graphs and the Sum-Product Algorithm. *IEEE T Inform Theory* **47**(2):498-519.

Lee, A. *et al.* (2010) On the Utility of Graphics Cards to Perform Massively Parallel Simulation of Advanced Monte Carlo Methods. *J. Comp. and Graph. Stat.* **19**(4):769-789.

Liang, F. *et al.* (2008) Mixtures of g Priors for Bayesian Variable Selection. *J. Am. Stat. Assoc.* **103**(481):410-423.

Lu, Y. *et al.* (2011) Kinome siRNA-phosphoproteomic screen identifies networks regulating Akt signaling. *Oncogene* **30**:4567-4577.

Maher, B. (2012) ENCODE: The human encyclopaedia. *Nature* **489**(7414):46-48.

Mukherjee, S, Speed, T.P. (2008) Network inference using informative priors. *Proc. Nat. Acad. Sci. USA* **105**(38):1431314318.

Mukherjee, S., Hill, S.M. (2011) Network clustering: probing biological heterogeneity by sparse graphical models. *Bioinformatics* **27**(7):994-1000.

Murphy, K. (2002) Dynamic Bayesian Networks: Representation, Inference and Learning, PhD Thesis, University of California, Berkeley.

Neve, R. *et al.* (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10**(6):515-527.

Oates, C., Mukherjee, S. (2012) Network Inference and Biological Dynamics. *Ann. Appl. Stat.* **6**(3):1209-1235.

Oates, C. *et al.* (2012b) Network Inference Using Steady-State Data and Goldbeter-Koshland Kinetics. *Bioinformatics* **28**(18):2342-2348.

Pearl, J. (1998) Why there is no statistical test for confounding, why many think there is, and why they are almost right. Technical report, Department of Statistics, UCLA, UC Los Angeles.

Pearl, J. (2009) Causal inference in statistics: An overview. *Stat. Surveys* **3**:96-146.

Rodríguez, A., Lenkoski, A., Dobra, A. (2011) Sparse covariance estimation in heterogeneous samples. *Electron. J. Statist.* **5**:981-1014.

Penfold *et al.* (2012) Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. *Bioinformatics* **28**(12):i233-i241.

Scott, J.G., Berger, J.O. (2010) Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *Ann. Stat.* **38**(5):2587-2619.

Smith, M. *et al.* (2001) Nonparametric regression using linear combinations of basis functions. *Stat. Comp.* **11**(4):313-322.

Song, L., Kolar, M., Xing, E.P. (2009) Time-Varying Dynamic Bayesian Networks. *Proceedings of the Conference on Advances in Neural Information Processing Systems (NIPS)* **22**:1732-1740.

Spencer, S., Hill, S.M., Mukherjee, S. (2012) Dynamic Bayesian networks for interventional data. *CRiSM Working Paper Series, The University of Warwick, UK* **12**:24.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population scale sequencing. *Nature* **467**(7319):1061-1073.

The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature* **490**(7418):61-70.

Vu, D., Hunter, D.R., Schweinberger, M. (2012) Model-based clustering of large networks. *Ann. Appl. Stat.* to appear.

Wei, P., Pan, W. (2012) Bayesian Joint Modeling of Multiple Gene Networks and Diverse Genomic Data to Identofy Target Genes of a Transcription Factor. *Ann. Appl. Stat.* **6**(1):334-355.

Weile, J. *et al.* (2012) Bayesian integration of networks without gold standards. *Bioinformatics* **28**(11):1495-1500.

Werhli, A.V., Husmeier, D. (2008) Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions. *Journal of Bioinformatics and Computational Biology* **6**(3):543-572.

Xu, T. *et al.* (2010) Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. *Sci. Sig.* **3**(113):ra20.

Yang, S. *et al.* (2012) Fused Multiple Graphical Lasso. arXiv:1209.2139 [cs.LG].

Zalatan, J.G. *et al.* (2012) Conformational Control of the Ste5 Scaffold Protein Insulates Against MAP Kinase Misactivation. *Science* **337**(6099):1218-1222.

Zellner, A. (1986) On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions, *Bayesian Inference and Decision Techniques - Essays in Honor of Bruno de Finetti, eds. P. K. Goel and A. Zellner*, 233-243.

Zhou, H., Pan, W., Shen, X. (2009) Penalized model-based clustering with unconstrained covariance matrices. *Electron. J. Stat.* **3**:1473-1496.